National Energy Research Scientific Computing Center (NERSC)

The Divergence Problem

Horst D. Simon
Director, NERSC Center Division, LBNL
February 24, 2003
http://www.nersc.gov/~simon

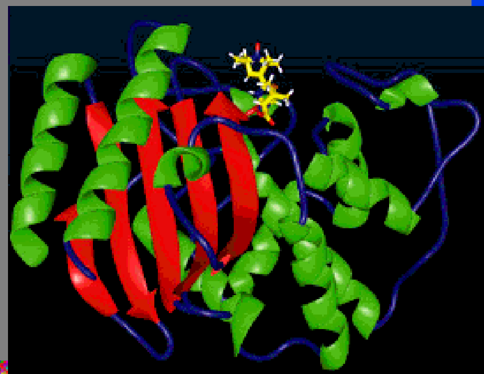
Outline

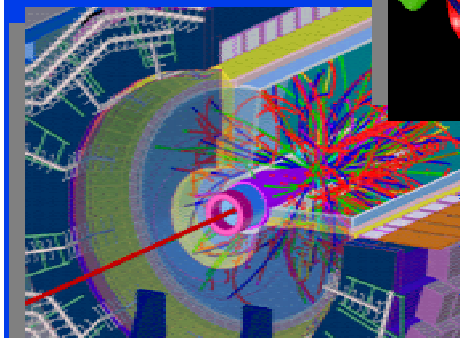
- Introducing NERSC
- Signposts of Change in 2002
- The Divergence Problem
- What should we do about it?

NERSC Center Overview

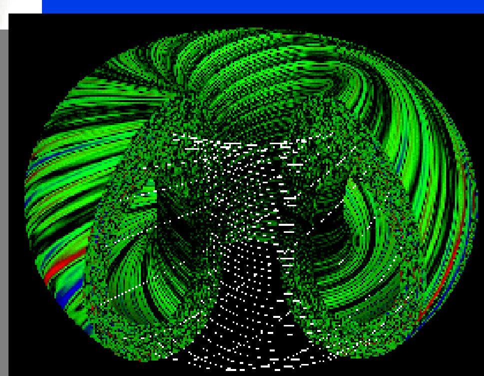
- Funded by DOE, annual budget $28M, about 65 staff
- Supports open, unclassified, basic research
- Located in the hills next to University of California, Berkeley campus
- close collaborations between university and NERSC in computer science and computational science
- close collaboration with about 125 scientists in the Computational Research Division at LBNL

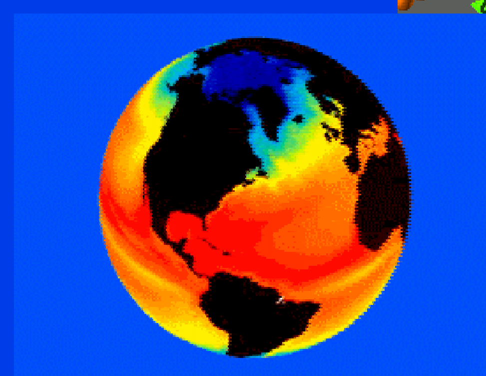
National Energy Research Scientific Computing Center





- Serves all disciplines of the DOE Office of Science
- ~2000 Users in ~400 projects
- Focus on large-scale computing





NERSC Goal: Enabling Scientific Discoveries

- Borrill (LBNL) + CalTech + others.

- BOOMERANG Experiments – analyze cosmic microwave background radiation data to obtain a better understanding of the universe
- The data analysis provides strong evidence that the geometry of the universe is flat
- Developed MADCAP software and provided computational capability on NERSC platforms.

Nature, April 27, 2000

Components of the Next-Generation NERSC

Outline

- Introducing NERSC
- Signposts of Change in 2002
- The Divergence Problem
- What should we do about it?

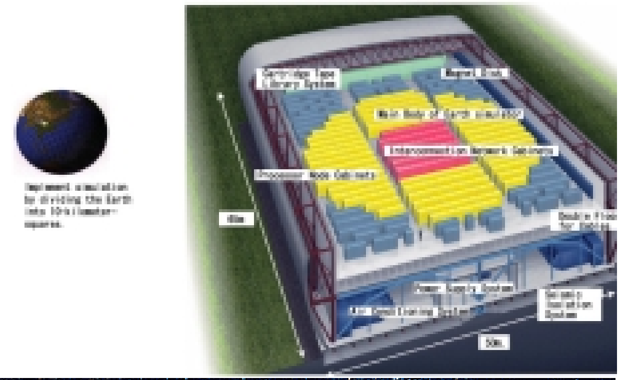
Signposts of Change in HPC

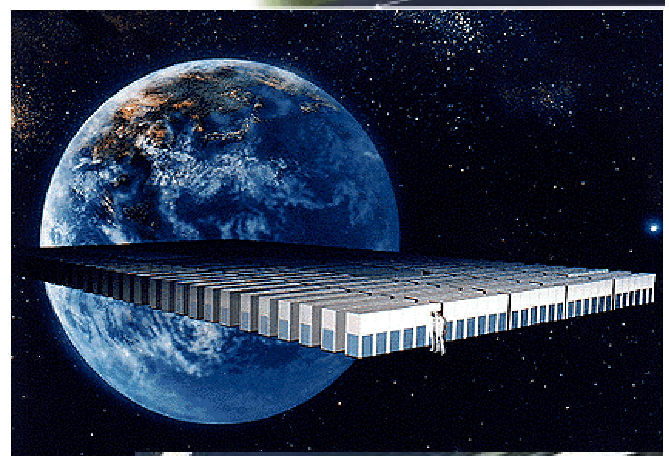
In early 2002 there were several signposts, which signal a fundamental change in HPC in the US:

- Installation and very impressive early performance results of the Earth Simulator System (April 2002)
- Lack of progress in computer architecture research evident at Petaflops Workshop (WIMPS, Feb. 2002)
- Poor or non-existing benchmarks on sustained systems performance (SSP) for the NERSC workload (March 2002)

The Earth Simulator System

- Based on the NEC SX architecture, 640 nodes, each node with 8 vector processors (8 Gflop/s peak per processor), 2 ns cycle time, 16GB shared memory.
  - Total of 5104 total processors, 40 TFlop/s peak, and 10 TB memory.
- It has a single stage crossbar (1800 miles of cable) 83,000 copper cables, 16 GB/s bandwidth, into and out of each node.
- 700 TB disk space
- 1.6 PB mass store
- 30,000 sqft computer room





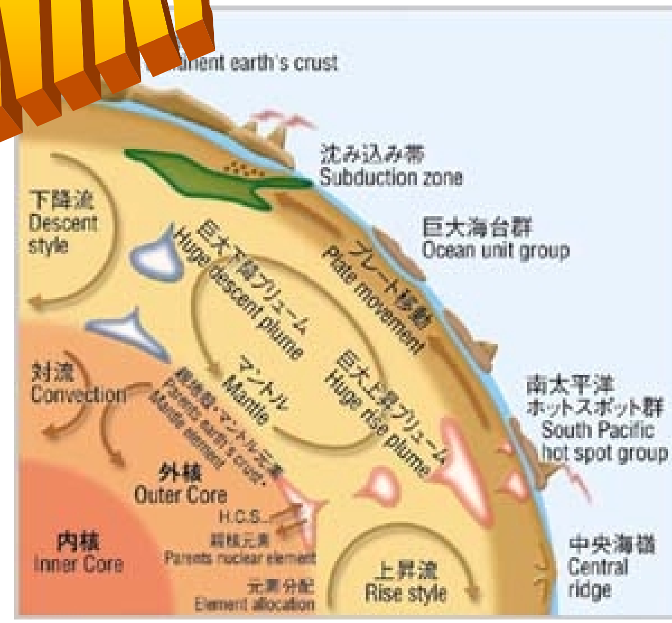
20th TOP500 List: The TOP10

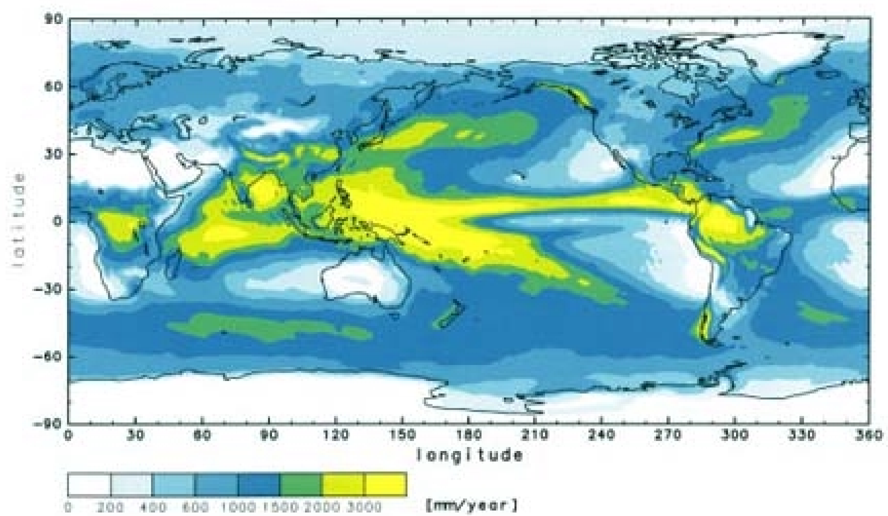
Rank	Manufacturer	Computer	R_{max} [TF/s]	Installation Site	Country	Year	Area of Installation	# Proc
1	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	Research	5120
2	HP	ASCI Q, AlphaServer SC	7.73	Los Alamos National Laboratory	USA	2002	Research	4096
2	HP	ASCI Q, AlphaServer SC	7.73	Los Alamos National Laboratory	USA	2002	Research	4096
4	IBM	ASCI White SP Power3	7.23	Lawrence Livermore National Laboratory	USA	2000	Research	8192
5	Linux NetworX	MCR Cluster	5.69	Lawrence Livermore National Laboratory	USA	2002	Research	8192
6	HP	AlphaServer SC ES45 1 GHz	4.46	Pittsburgh Supercomputing Center	USA	2001	Academic	3016
7	HP	AlphaServer SC ES45 1 GHz	3.98	Commissariat a l'Energie Atomique (CEA)	France	2001	Research	2560
8	HPTi	Xeon Cluster - Myrinet2000	3.34	Forecast Systems Laboratory - NOAA	USA	2002	Research	1536
9	IBM	pSeries 690 Turbo	3.16	HPCx	UK	2002	Academic	1280
10	IBM	pSeries 690 Turbo	3.16	NCAR (National Center for Atmospheric Research)	USA	2002	Research	1216

The Earth Simulator i… an

COMPUTENIK!

- Linpack benchn… TF/s = 87% of 4(…
- Completed Apr…
- Driven by clima… …d earthquake simulation
- Gordon Bell Prize at SC2002





http://www.es.jamstec.go.jp/esrdc/eng/menu.html

Understanding and Prediction of Global Climate Change	Understanding of Plate Tectonics
Occurrence prediction of meteorological disaster	Understanding of long-range crustal movements
Occurrence prediction of El Niño	Understanding of mechanism of seismicity
Understanding of effect of global warming	Understanding of migration of underground water and materials transfer in strata
Establishment of simulation technology with 1km resolution	

Catalyst for fundamental change in U.S. science policy or call for a small course correction?

- The important event is not a single machine but the commitment of the Japanese government to invest in science-driven computing.
- U.S. computer industry is driven by commercial applications -- not focused on scientific computing.
- The Earth Simulator is a direct investment in scientific computing, giving Japanese scientific communities a material advantage and making them more attractive as international collaborators.
- **The Earth Simulator is not a special purpose machine:** *All* U.S. scientific computing communities are potentially now at a handicap of 10 to 100 in delivered computing capability.

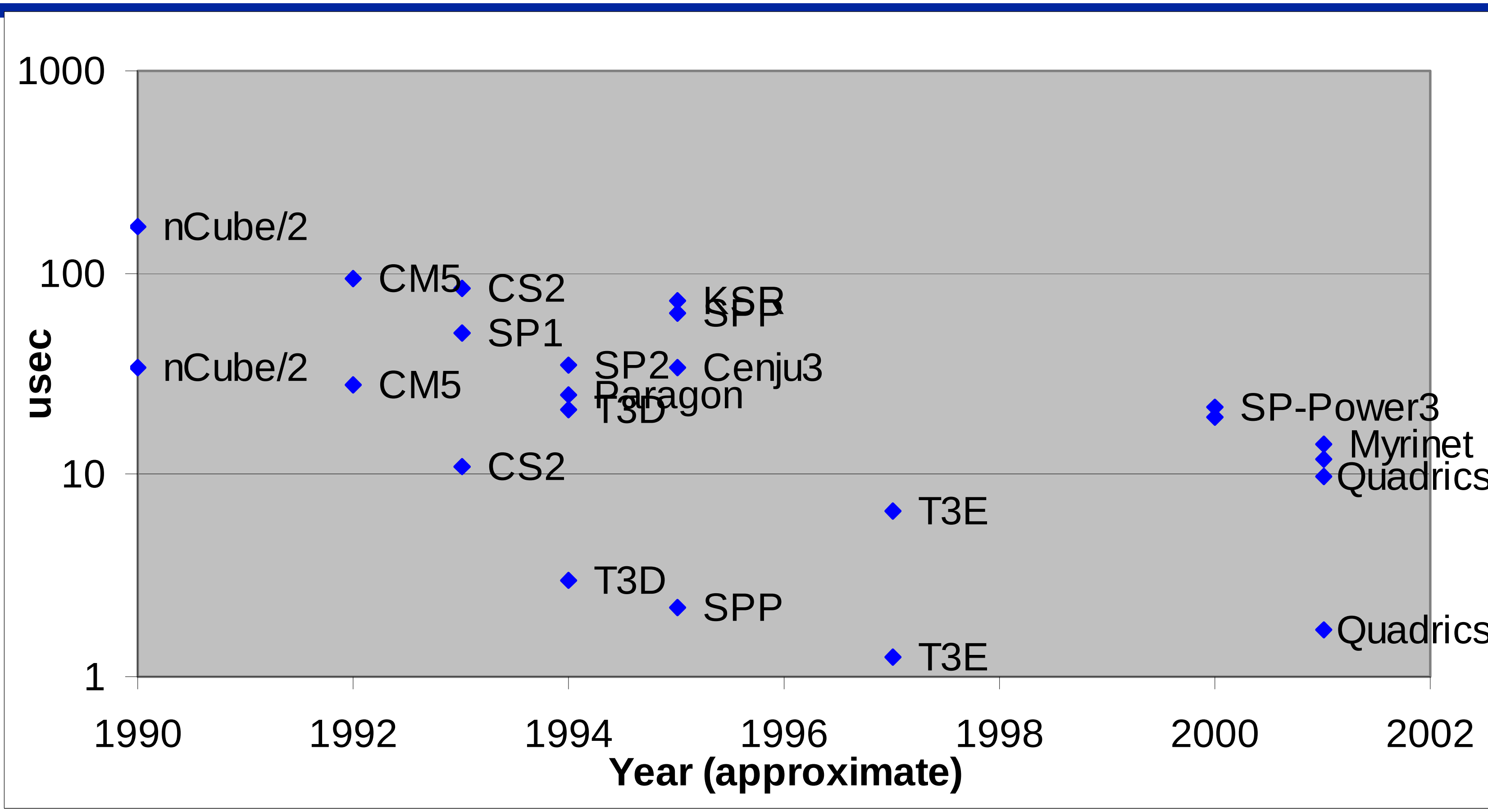
Perspective

- Peak performance does not reveal the real impact of the Earth Simulator.
- Japanese scientific policy is to build strategic partnerships in climate, nanoscience and fusion, moving to dominate simulation and modeling in many disciplines – not just climate modeling.
- To optimize architectures for scientific computing, it is necessary to establish the feedback between scientific applications and computer design over multiple generations of machines.
- The Japanese Earth Simulator project implemented one cycle of that feedback, and made dramatic progress.

Basic Research Issues/Observations

- WIMPS2002 = Petaflops 1997
  - no significant progress in five years
- Only a handful of supercomputing relevant computer architecture projects at US universities; versus of the order of 50 in early 1990s
- Lack of interest in supporting supercomputing relevant basic research
  - parallel language and tools research has been almost abandoned
  - focus on grid middleware and tools

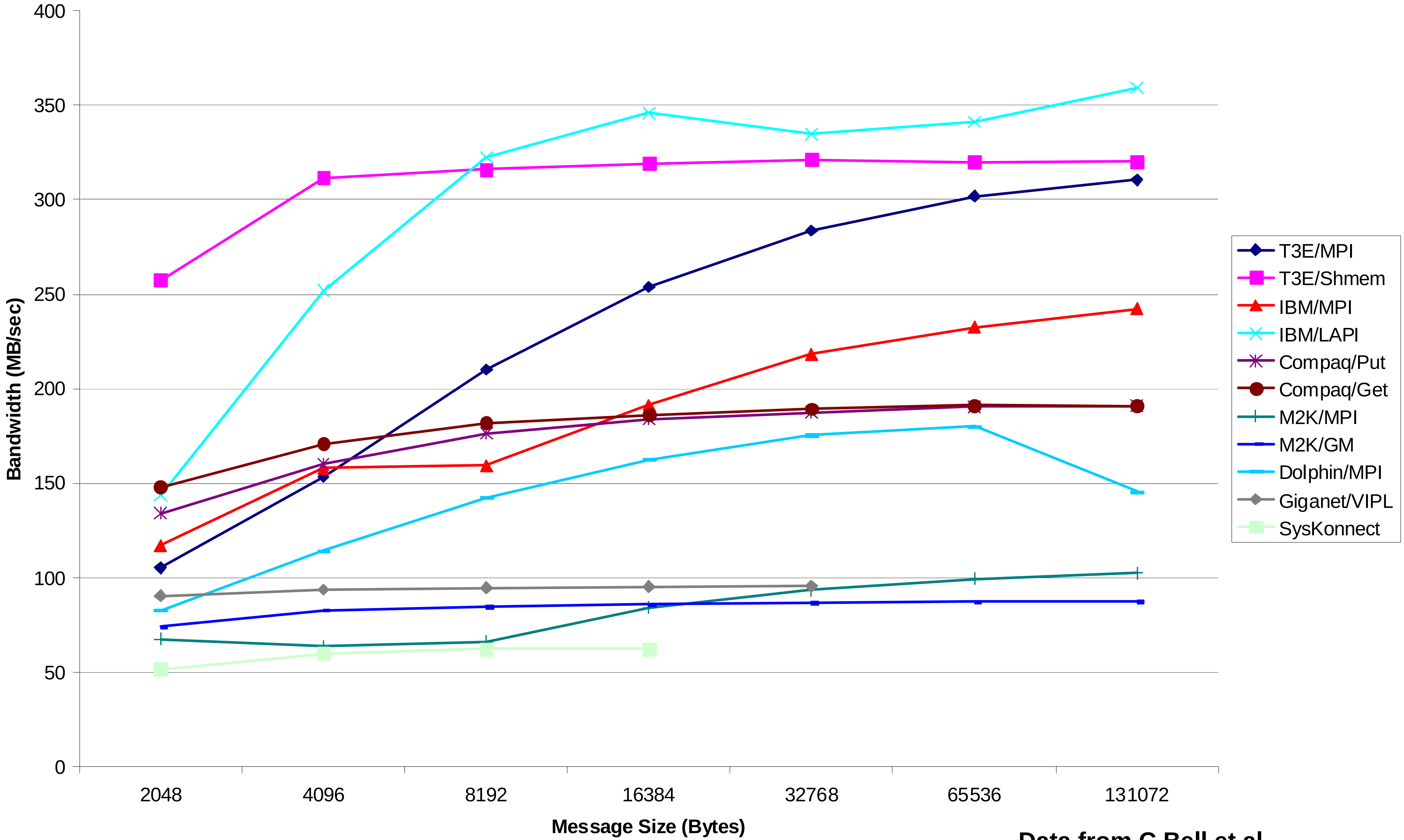
End to End Latency Over Time



- Latency has not improved significantly
  - T3E (shmem) was lowest point
  - Federation in 2003 will not reach that level – 7 years later!

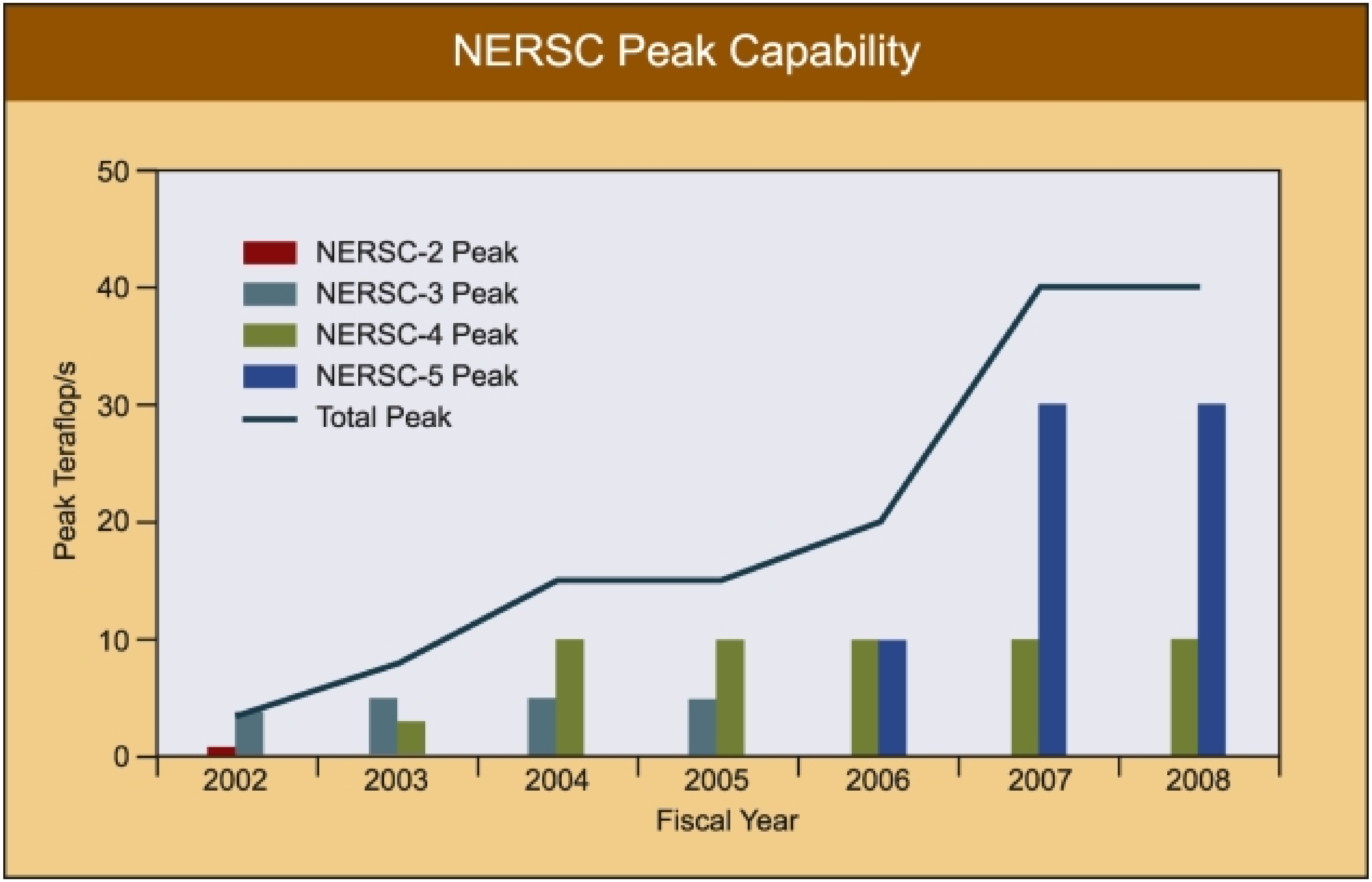
Data from C. Bell et al. "An Evaluation of Current High-Performance Networks" see http://upc.nersc.gov/publications

Bandwidth Chart



Data from C.Bell et al.

NERSC Peak Capability as projected in the Strategic Plan



Combined NERSC-3 Characteristics

- The combined NERSC-3/4 system (NERSC-3Base and NERSC-3Enhanced) will have
  - 416 16 way Power 3+ nodes with each CPU at 1.5 Gflop/s
    - 380 for computation
  - 6,656 CPUs – 6,080 for computation
  - Total Peak Performance of 10 Teraflop/s
  - Total Aggregate Memory is 7.8 TB
  - Total GPFS disk will be 44 TB
    - Local system disk is an additional 15 TB
  - Combined SSP-2 measure is 1.238 Tflop/s
  - NERSC-3E in production by the end of Q1/CY03
    - Nodes will arrive in the first two weeks of November
    - 30 day availability test ends on 2/25/03

Comparison with Other Systems

	NERSC-3 E	ASCI White	ES	Cheetah (ORNL)	PNNL Mid 2003
Nodes	416	512	640	27	700
CPUs	6,656	8,192	5,120	864	1400
Peak(Tflops)	10	12	40	4.5	9.6(8.3)
Memory (TB)	7.8	4	10	1	1.8
Disk(TB)	60	150	700	9	53
SSP(Gflop/s)	1,238	1,652		179	

PNNL system available in Q3 CY2003

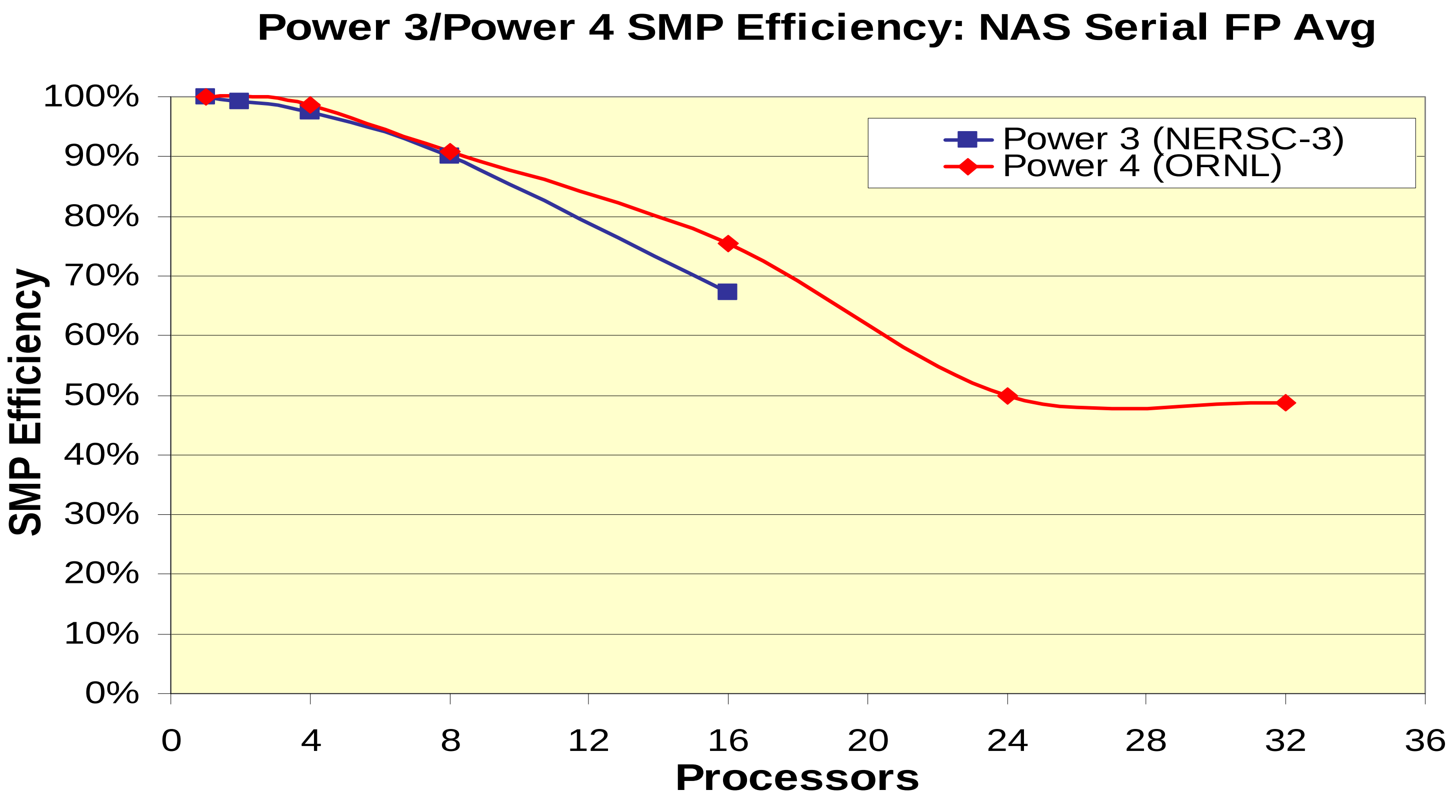
SSP = sustained systems performance (NERSC applications benchmark)

Power 4 in the NERSC Applications Benchmark

- The NERSC – 3 base system delivers

  618 Gflop/s on NERSC SSP
- We measured **179 Gflop/s** on the 4.5 Tflop/s peak Power 4 system at ORNL
- Assume a Power 4 system with same base cost as NERSC-3:
  - Available to NERSC users only in mid to late 2004
  - Only a 7% performance improvement 3 years after NERSC-3
- The performance of Power 4 is a clear indication of the **DIVERGENCE PROBLEM**
- Power 4 was not designed for scientific applications

Power 4 does not perform as well as expected



Power 4 versus Power 3

- By simple measures a Power 4+/Federation should be 4 to 10 times better than an equal number of Power 3 CPUs
  - 4.5 time the Gflop/s per CPU, 9 times the GFlop/s per node, 8 times the interconnect bandwidth, 11 times the memory bandwidth, etc
- Measured performance did not track with peak improvements
  - Average improvement for real applications was only 2.5 times better
  - The integrated SSP was actually worst than on Power 3
    - Few CPUs for the same cost

Why?

- Memory latency did not improve, in fact it got relatively worse.
  - Aggravated by the lack of rename registers that generated more flushes of the instruction pipeline
- Power-4 nodes do not scale well for more than 16 scientific tasks

Outline

- Introducing NERSC
- Signposts of Change in 2002
- The Divergence Problem
- What should we do about it?

Signposts of Change in HPC

In early 2002 there were several signposts, which signal a fundamental change in HPC in the US:

- Installation and very impressive early performance results of the Earth Simulator System (April 2002)
- Lack of progress in computer architecture research evident at Petaflops Workshop (WIMPS, Feb. 2002)
- Poor or non-existing benchmarks on SSP for the NERSC workload (March 2002)

This is happening against the backdrop of:

- increasing lack of interest in HPC by some US vendors
- further consolidation and reduction of the number of vendors (Compaq + HP merger)
- reduced profitability and reduced technology investments (dot com bust)

We are in the middle of a fundamental change of the basic premises of the HPC market in the U.S.

We have pursued the logical extreme of the "commodity parts" path.

Low cost path

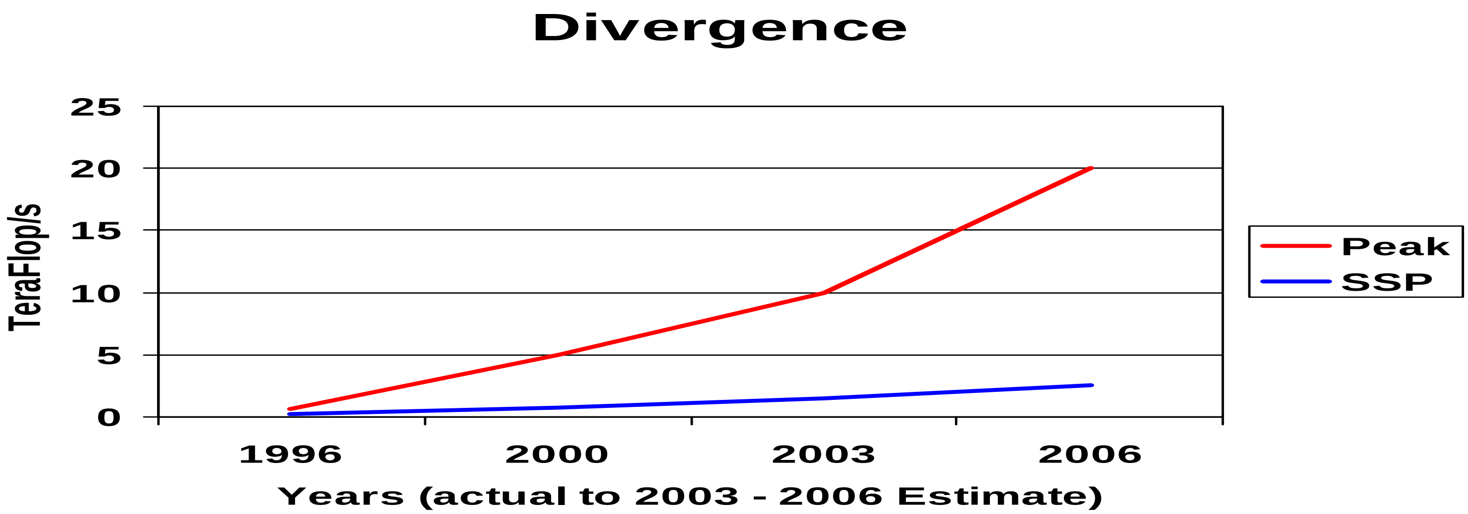
This Became

Clusters of Symmetric Multiprocessors:
Ensembles of Data Servers + Fast Switch

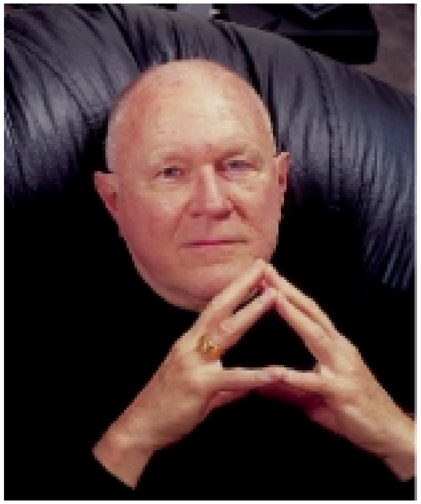
- The commodity building block was the microprocessor but is now the entire server (SMP).
- Communications and memory bandwidth are not scaling with processor power.
- We have arrived at near football-field size computers consuming megawatts of electricity.

The Divergence Problem

- The requirements of high performance computing for science and engineering and the requirements of the commercial market are diverging.
- The commercial cluster of SMP approach is no longer sufficient to provide the highest level of performance
  - Lack of memory bandwidth
  - High interconnect latency
  - Lack of interconnect bandwidth
  - Lack of high performance parallel I/O
  - High cost of ownership for large scale systems



Recent opinions on commodity technology in supercomputing



- "Gordon Bell, now a senior researcher at Microsoft, warns that off-the-shelf supercomputing is a dead end."

  quoted from MIT Technology Review, Feb 2003.

- "Beowulf is dead"

  Thomas Sterling, Caltech, quoted from a panel discussion at SC2002, Nov. 2002

The State of the American Computer Industry – In Scientific Computing

- The major players that are still active in scientific supercomputing are
  - IBM
  - Hewlett Packard
  - Cray (a small surviving and evolved portion)
  - Sun
  - SGI
- We don't have a building block optimized for scientific computation.
- The target commercial market is data and web serving, and that market dominates the economics of the computer industry beyond the personal computer.
- The architectural barriers for scientific computing stem from this situation
  - Memory bandwidth and latency (optimized for databases)
  - Interconnect bandwidth and latency (optimized for transaction processing)
- If you don't have a viable market for those building blocks, then how do you cause them to be created?

The Dead Supercomputer Society

- See http://www.paralogos.com/DeadSuper/
- list of 42 dead companies or projects from 1975 - today

Gone, But Not Forgotten: Evidence of Enormous Creativity in Computing in the U.S. ca. 1990

- ACRI
- Alliant
- American Supercomputer
- Ametek
- Applied Dynamics
- Astronautics
- BBN
- CDC
- Convex
- Cray Computer
- Cray Research
- Culler-Harris
- Culler Scientific
- Cydrome
- Dana/Ardent/Stellar/Stardent
- Denelcor
- Elexsi
- ETA Systems
- Evans and Sutherland Computer
- Floating Point Systems
- Galaxy YH-1
- Goodyear Aerospace MPP
- Gould NPL
- Guiltech
- Intel Scientific Computers
- International Parallel Machines
- Kendall Square Research
- Key Computer Laboratories
- MasPar
- Meiko
- Multiflow
- Myrias
- Numerix
- Prisma
- Tera
- Thinking Machines
- Saxpy
- Scientific Computer Systems (SCS)
- Soviet Supercomputers
- Supertek
- Supercomputer Systems
- Suprenum
- Vitesse Electronics

But this is not 1990

- Starting a number of new small companies seeded by federal research investment (as DARPA did in the HPCCI) is probably not feasible .
- There is now a much larger commercial market, and the industry dynamics are different.
- The Earth Simulator “event” has motivated IBM and others to better address the needs of the scientific community.
- There is still a significant scientific market for high performance computing outside of supercomputer centers.
- For this new environment, we need a new, sustainable strategy for the future of scientific computing.

Outline

- Introducing NERSC
- Signposts of Change in 2002
- The Divergence Problem
- What should we do about it?

Need for a Sustainable Effort

- Without a sustained effort, scientific communities cannot invest their efforts and resources to adapt their computing strategy to new classes of hardware.
- Parallel computing itself required a decade to find scalable algorithms to make it useful, and the process is still continuing.
- The U.S. policy should not be to create one machine just to show we can do it, but should be a long-term program that ensures preeminence in scientific computing.
- The most powerful of these systems need to be available to the open, scientific community (in addition to any special communities)

Why Does Cost Matter?

If this is so important, why does cost matter?

- If effective scientific supercomputing is only available at high cost, it will have impact on only a small part of the scientific community.
- So, need to leverage the resources of mainstream IT companies like IBM, HP and Intel as well as any special architecture companies like Cray.
- And the national science policy should motivate them to participate durably.

Creating a New Class of Computer Architectures for Scientific Computing

- Sustained cooperative development of new computer architectures
- A focus on sustained performance of scientific applications – not on peak performance!
- Addressing the key bottlenecks of bandwidth and latency for memory and processor interconnection
- A strategy to pursue several architectures at multiple sites
- A new investment in the computer science research and scientific research communities

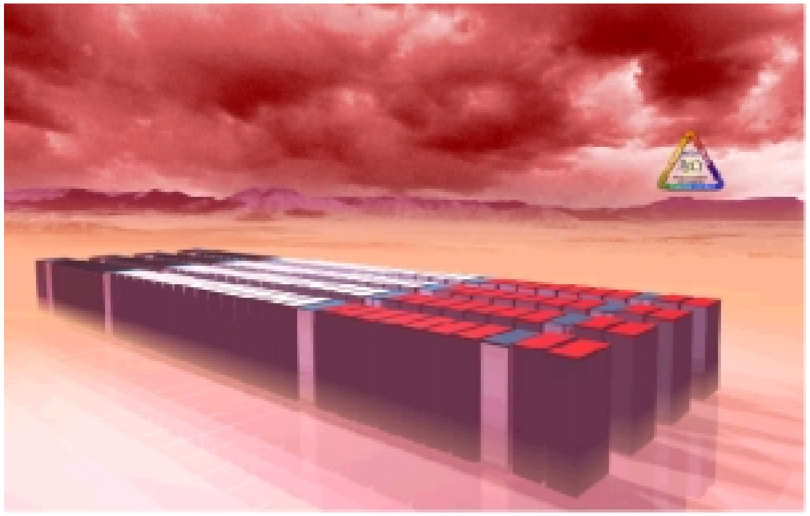
A New Architecture Strategy: Beyond Evaluation to Cooperative Development

A proposal to establish feedback between science and computer design lasting for generations of machines

- Application teams to drive the design of new architectures
- Continued, simultaneous evaluation of multiple scientific applications replacing "rules of thumb" for computer designers
  - Example is the Performance Evaluation Research Center (PERC)
- Leveraging current components and research prototypes into new architectures
- Continual redesign and testing of prototypes in a vendor partnership to create new scientific computers
- Addressing the scientific market beyond lab and academic supercomputer centers

First Example of Cooperative Development: "Red Storm"

- Collaboration between Sandia Natl. Lab. and Cray
- True MPP, designed to be a single system
- Distributed memory MIMD parallel supercomputer
- Fully connected 3-D mesh interconnect. Each compute node processor has a bi-directional connection to the primary communication network.
- 108 compute node cabinets and 10,368 compute node processors (AMD Sledgehammer @ 2.0 GHz) ~20 Tflop/s peak
- ~10 TB of DDR memory @ 333 MHz
- 240 TB of disk storage (120 TB per color)
- Less than 2 MW total power and cooling.
- Less than 3,000 square feet of floor space



Courtesy: Bill Camp and Jim Thompkins, Sandia

Cooperative Development – NERSC/ANL/IBM Workshop

• Goal: Pursue a path(s) to provide a system that can have sustained performance in the range of 30-50% on systems with peak performances of more than one petaflop/s....

•Shorter term goal: By 2005, field a computer at twice the applications performance of the Earth Simulator that is on a sustainable path for scientific computing

•Held two joint workshops

- •Sept 2002 – defining the Blue Planet architecture
- •Nov. 2002 – IBM gathered input for Power 6

•Developed White Paper "Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership,“ available at http://www.nersc.gov/news/blueplanet.html

Selection is Based on Scientific Applications

	AMR	Coupled Climate	Astrophysics		Nanoscience	
			MADCAP	Cactus	FLAPW	LSMS
Sensitive to global bisection	X	X	X		X	
Sensitive to processor to memory latency	X	X			X	
Sensitive to network latency	X	X	X	X	X	
Sensitive to point to point communications	X	X				X
Sensitive to OS interference in frequent barriers				X	X	
Benefits from deep CPU pipelining	X	X	X	X	X	X
Benefits from Large SMP nodes	X					

“Blue Planet”: Extending IBM Power Technology and “Virtual Vector” Processing

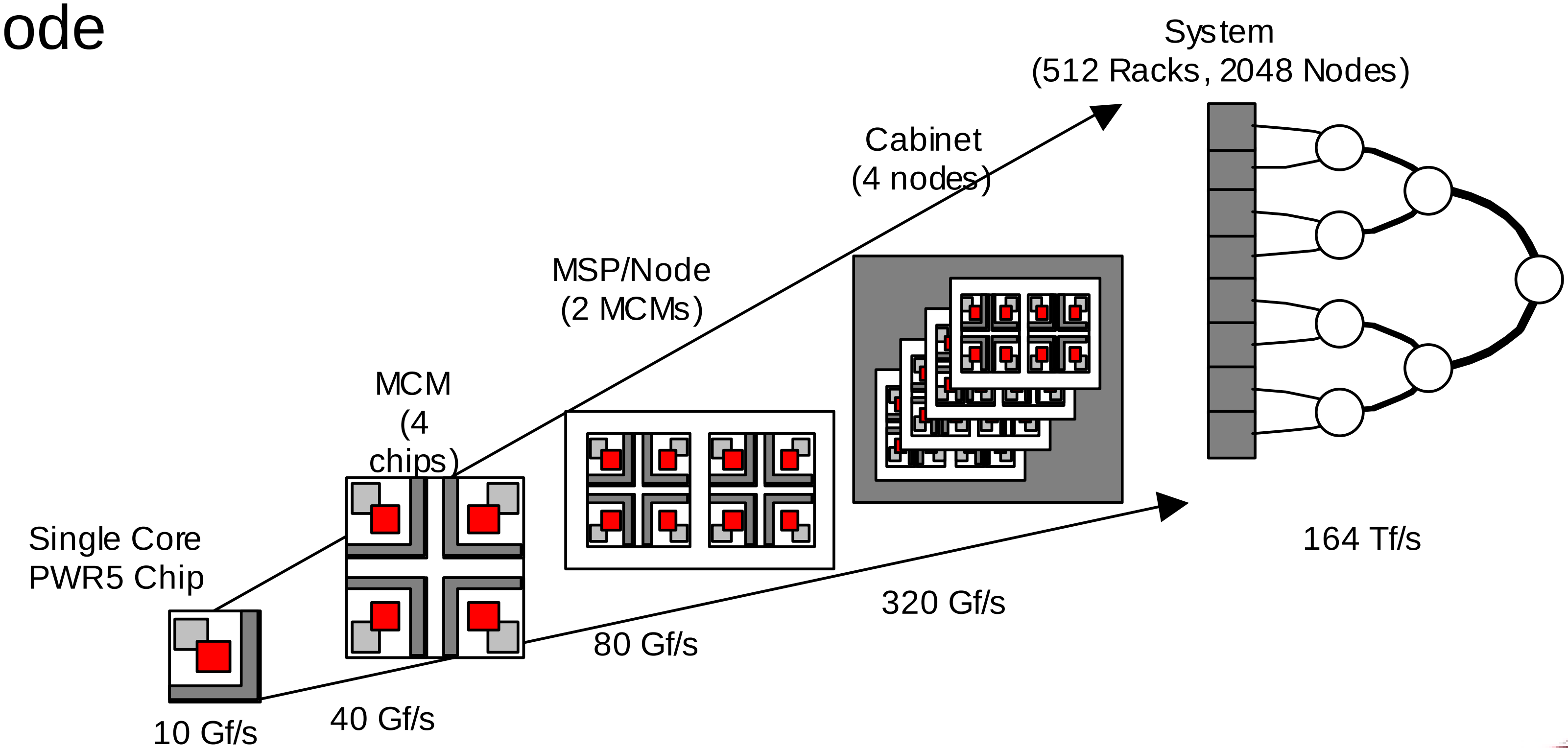
Addressing the key barriers to effective scientific computing

- — Memory bandwidth and latency
- — Interconnect bandwidth and latency
- — Programmability for scientific applications

- The Strategy is to get back“inside the box” of commercial servers (SMPs)
  - — Increasing memory and switch bandwidth using commercial parts available over the the next two years
- Exploration of new architectures with the IBM design team
- Enabling the vector programming model inside a Power 5 SMP node
- Changing the design of subsequent generations of microprocessors

Blue Planet: A Conceptual View

- Increasing memory bandwidth – single core chips with dedicated caches for 8 way nodes
- Increasing switch bandwidth and decreasing latency
- Enabling “vector” programming model inside each SMP node

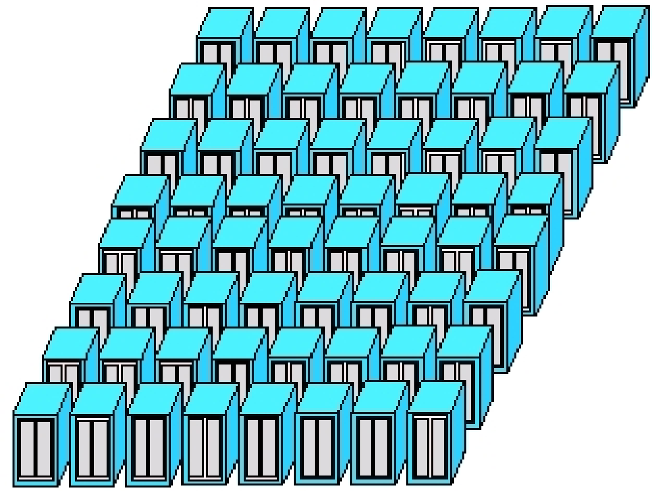


Ultracomputer Research:

ibm.com/eserver

Blue Planet

System
(256 racks/
2,048 nodes/
16,384 processors
+ 160 switch frames)



160 TF/s

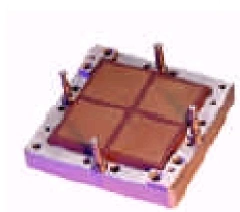
Rack
(64 processors/
8 nodes)

640 GF/s

ViVA Node
(8 processors)

80 GF/s

MCM
(4 processors)



40 GF/s

POWER5+ Chip
(1 processor)

10 GF/s

Blue Planet Target Design:

- ✓ POWER5+ GS single-core chip
- ✓ Approx 2.5 GHz
- ✓ 0.10u 10S2 technology
- ✓ 2005 availability

http://www.nersc.gov/news/blueplanetmore.html

Slide courtesy of
Peter Ungaro, IBM

Why this is not Business as Usual for IBM

- Introducing 8 way Power 5+ nodes with single cores early is entirely new packaging
  - For power 4, 8 way nodes came out 18 months after full size SMP (32 CPUs)
  - Each CPU will have its own L1, L2 and L3 cache
  - Each node will have twice the number of memory buses as standard nodes
  - 8 way nodes will run at full clock rate (as opposed to the slower dual core 8 way nodes soon to be introduced.
- Synchronizing CPUs (“Virtual Vectors) is not in their plan
  - Both hardware and compiler technology involved
- An additional stage (level) in the Federation switch is not in their plan
  - Increases a factor of 4 in number of links.
- Decreasing switch latency is not in their plan
  - Requires a radical redesign of their software stack
- Operating System, Compiler, Library and Scalability Improvements

Managing Long-Term Architecture Development

- DOE Lab system is ideally suited to manage large-scale, long-term research and development
- We believe that long-term participation from the universities is critical to the success of this proposed initiative
- We need to engage architects, scientists, computer scientists in a way that is accountable to one agency
  - And to do that over multiple generations
  - And with multiple vendors
- These have to be run as closed-loop integrated projects
- We need to avoid the past failure modes of interagency development

Conclusion

- We have pursued the logical extreme of the "commodity parts" path.
- This path was a cost-efficient "free ride" on a Moore's Law growth curve
- The divergence problem shows that this free ride is coming to an end.
- Business as usual will not preserve U.S. leadership in advanced scientific computing
- New computer architectures optimized for scientific computing are critical to enable 21st Century Science
- The HPC center and user community needs to develop these in a new mode of sustainable partnership with the vendors

U.S. science requires a strategy to create cost-effective, science-driven computer architectures.